A TWO-LOCUS NEUTRALITY TEST: APPLICATIONS TO HUMANS, *E. COLI* AND LODGEPOLE PINE

PHILIP W. HEDRICK[1] AND GLENYS THOMSON

Department of Genetics, University of California, Berkeley, California 94720

Manuscript received July 3, 1984
Revised copy accepted September 19, 1985

ABSTRACT

The expected disequilibrium between two loci with k alleles at one locus and l alleles at the other is given for a sample of size n drawn from a population under neutrality equilibrium. Three different measures of disequilibrium with 95% intervals are tabulated for combinations of n, k, l and $4Nc$, where N is the effective population size and c is the amount of recombination between the loci. The extent and pattern of disequilibrium are strongly dependent upon $4Nc$ and are somewhat dependent on n, k and l. The 95% intervals are large, particularly for low numbers of alleles and low values of $4Nc$. As examples, observed disequilibrium from histocompatibility loci in humans (HLA) and electrophoretic data in *E. coli* and lodgepole pine were compared to these theoretical values. Using information about recombination rates, the HLA data showed more disequilibrium than neutrality expectations, whereas electrophoretic data from *E. coli* and lodgepole pine had somewhat less disequilibrium than neutrality expectations.

THE neutrality theory assumes that all alleles at a locus are selectively equivalent, that mutation is the source of new genetic variation and that genetic drift results in the loss of variation. Given these conditions, a population at equilibrium under neutrality has an expected distribution of alleles for each locus (*e.g.*, KIMURA and OHTA 1971). In addition, a population at equilibrium under neutrality has an expected association between alleles at linked loci. Even though there is no selection among different alleles at a locus, the combined effects of mutation and genetic drift result in an interlocus association when there is limited recombination (*e.g.*, OHTA and KIMURA 1969; HILL 1975).

When considering variation at a single locus, a number of approaches have been suggested to determine whether the pattern of genetic variation is consistent with neutrality (see EWENS 1977 for a review). One particularly ingenious approach was developed by EWENS (1972), in which he gave the distribution of alleles expected in a sample of a given size obtained from a population at equilibrium under neutrality. Extending these results, WATTERSON (1978a,b) gave a statistical test that allows the comparison of the observed

[1] Permanent address: Division of Biological Sciences, University of Kansas, Lawrence, Kansas 66045.

Hardy-Weinberg homozygosity in a sample of a given size that contains k different alleles with the homozygosity expected in a sample of the same size with the same number of alleles drawn from a population at equilibrium under neutrality.

A general approach for consideration of the expected properties of gametic samples of closely linked loci is particularly relevant now because of the rapidly accumulating data on gametic frequencies of closely linked molecular variants identified either by restriction endonucleases or nucleotide sequencing (W. KLITZ and P. W. HEDRICK, unpublished results). A number of studies have investigated such related topics as the genetic variation in a multiallelic, two-locus neutrality population (*e.g.*, HILL 1975; TAKAHATA 1982) and the distribution of genetic variation in samples from a two-allele, two-locus neutrality population (HILL 1981; GOLDING 1984). However, the extension of EWENS' sampling approach to two loci with multiple alleles appears to be mathematically very difficult. As a result, we have determined the properties of two-locus samples using computer simulations. The simulation technique is that developed by HUDSON (1983) and described in detail in his manuscript. Here we shall examine the two-locus association in a sample of size n genes (where $n/2$ is the number of diploid individuals) with k and l different alleles at the two different loci. These stimulation results are then compared to the observed associations for genetic variants at histocompatibility loci in humans and electrophoretic loci from *Escherichia coli* and lodgepole pine.

MEASURES OF GAMETIC DISEQUILIBRIUM

The extent of association between alleles at different loci (gametic disequilibrium) can be measured in several ways for a specific gamete (see HEDRICK, JAIN and HOLDEN 1978 for a review). A widely used measure of gametic disequilibrium for a given gamete is

$$D_{ij} = x_{ij} - p_i q_j \tag{1}$$

where x_{ij} is the observed frequency of gamete A_iB_j, p_i and q_j are the frequencies of alleles A_i and B_j at loci A and B, and the expected frequency of gamete A_iB_j is p_iq_j, assuming no association between the alleles. The range of this measure of gametic disequilibrium is a function of the allelic frequencies. Therefore, a measure that has the same range for all allelic frequencies is desirable. For this reason, LEWONTIN (1964) suggested using the measure

$$D_{ij}' = \frac{D_{ij}}{D_{\max}} \tag{2}$$

where if $D_{ij} < 0$, $D_{\max}$ is the lesser of p_iq_j and $(1 - p_i)(1 - q_j)$ and if $D_{ij} > 0$, $D_{\max}$ is the lesser of $p_i(1 - q_j)$ and $(1 - p_i)q_j$.

The total disequilibrium between all the alleles at two loci can be measured in several ways. For example, the total disequilibrium can be expressed as

$$D^2 = \sum_{i=1}^{k} \sum_{j=1}^{l} D_{ij}^2. \tag{3}$$

However, this measure, like D_{ij}, is highly dependent upon allelic frequencies. As a result, other measures of association, such as those given in expressions (5), (7) and (9), that are not so dependent on allelic frequencies are preferable.

One approach to defining a two-locus, multiallelic measure that is not dependent upon allelic frequencies is to standardize the measure by the single-locus heterozygosity. For example, the Hardy-Weinberg homozygosity at locus A having k alleles is

$$F_A = \sum_{i=1}^{k} p_i^2 \tag{4a}$$

and for locus B with l alleles is

$$F_B = \sum_{j=1}^{l} q_j^2. \tag{4b}$$

Using the complements of these expressions, the Hardy-Weinberg heterozygosities (or genetic diversities for haploid genomes), a standardized measure of two-locus association is

$$D^* = \frac{D^2}{(1 - F_A)(1 - F_B)} \tag{5}$$

a measure termed R by MARUYAMA (1982) and similar to σ_D^2 as given by HILL (1975) and OHTA (1980) (see the discussion in the RESULTS section). If we define

$$F_{AB} = \sum_{i=1}^{k} \sum_{j=1}^{l} x_{ij}^2 \tag{6}$$

then another measure of association is

$$F^* = F_{AB} - F_A F_B \tag{7}$$

called the identity excess by OHTA (1980). In addition, the statistic

$$Q = n \sum_{i=1}^{k} \sum_{j=1}^{l} \frac{D_{ij}^2}{p_i q_j} \tag{8}$$

gives another standardized measure of total disequilibrium. Q is approximately χ^2 distributed under the null hypothesis of $D_{ij} = 0$ (see HILL 1975) with $(k - 1)(l - 1)$ degrees of freedom. If there is no association between alleles at the loci, then Q divided by the degrees of freedom should equal unity within sampling fluctuations. In order to make this measure sample size independent (BISHOP, FEINBERG and HOLLAND 1975) when $D \neq 0$, the measure

$$Q^* = \frac{Q}{n(k - 1)(l - 1)} \tag{9}$$

is useful.

RESULTS

In this section, we shall give the values calculated from simulations for various measures of two-locus associations in a sample of size n with k and l different alleles at loci A and B, respectively. These measures of association are dependent upon the amount of recombination as measured by the quantity $4Nc$. N is the size of the population from which the sample is drawn, and c is the rate of recombination between the loci. When $4Nc = 0$, there is no recombination between the loci. If $4Nc \gg 1$, then the recombination ratio should be large enough that there would be little association of alleles at different loci.

The simulated samples using the program of HUDSON (1983) were drawn from an equilibrium population. As a result, allelic frequencies for a given k and l are independent of $4Nu$, where u is the mutation rate to a new allele (EWENS 1972). Simulations with different $4Nu$ values were run for each $4Nc$ and n combination to obtain samples with a wider range of k and l values. From extensive simulations, HUDSON (1985) has shown that the disequilibrium values conditioned on the number of alleles are nearly independent of $4Nu$. For each combination of parameters, between 400 and 2500 replicate samples were generated in order to obtain 95% empirical confidence limits, termed 95% intervals herein, for the three measures given by expressions (5), (7) and (9).

When the recombination rate is low, some of the two-locus gametes possible are low in frequency or are missing. In fact, even when $4Nc$ is large, some gametes are still missing from the sample. If there are k and l alleles at the two loci, then there are kl possible two-locus gametes. With no recombination, new gametes are only produced by mutation so that a relatively small proportion of the gametes possible would be expected. However, as the recombination rate increases, then the number of gametic types present in the sample increases. To illustrate this effect, Figure 1 gives the proportion of possible gametes actually observed for three different numbers of alleles for $n = 200$. For example, if there were eight alleles at both loci and no recombination, then only about one-fifth of the possible gametes are observed. When there is a high level of recombination, $4Nc \geqq 100$, then approximately twice as many different gametic types are observed. The values for sample sizes of $n = 100$ and 400 are nearly identical to these, although there is a slight increase in the proportion of possible gametes observed as the sample size increases.

The distribution of D_{ij} values for a given number of alleles at both loci is nearly symmetrically distributed around zero, with variation that increases as $4Nc$ becomes lower. Like the distribution of D_{ij} values, the distribution of D'_{ij} values, for those D'_{ij} values that are not 1.0 or -1.0, has a mean near zero (actually slightly greater than zero) and a nearly symmetrical distribution. Figures 2 and 3 give the distribution of D'_{ij} values for examples in which $4Nc$ is large ($4Nc = 500$) and small ($4Nc = 10$), respectively. (Note that the vertical scales in these figures are different.)

The proportion of D'_{ij} values that are either -1.0 or 1.0, *i.e.*, those that are maximally associated, is greatly affected by recombination. As an example,

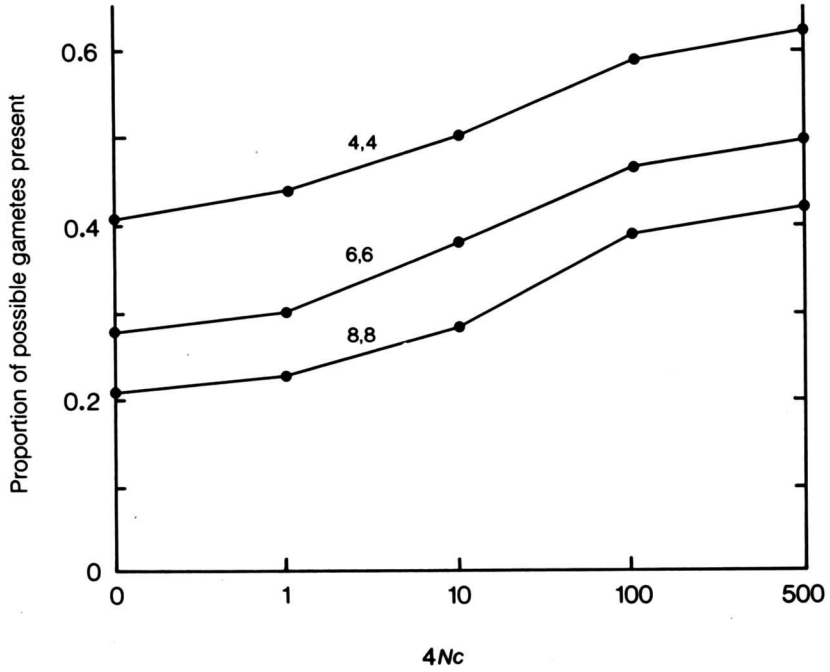

FIGURE 1.—Proportion of possible gametes present in samples of size n = 200 when there are 4, 6 and 8 alleles at the two loci for different levels of recombination.

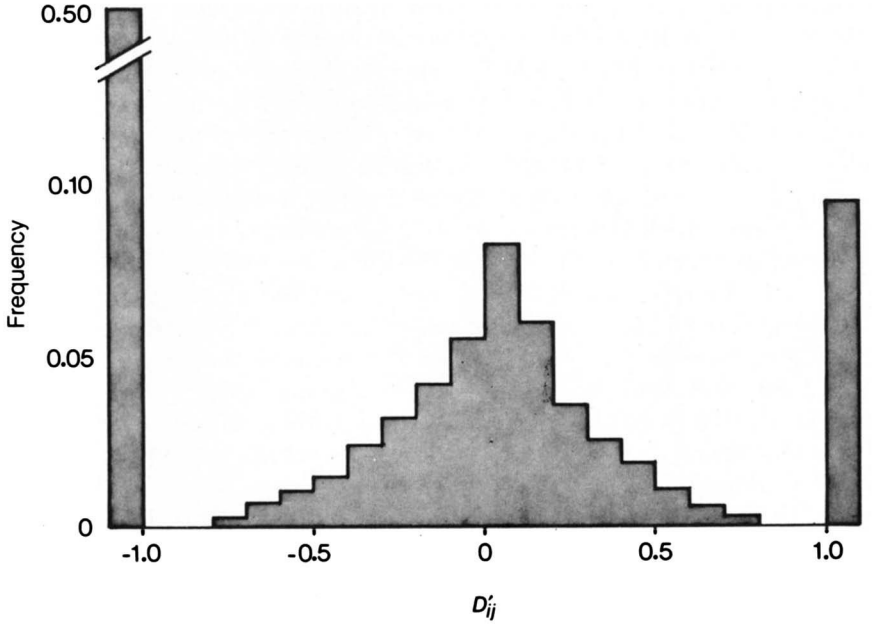

Figure 2.—The distribution of D'_{ij} in a sample of size n = 200 and $4Nc$ = 500 when there are six alleles at both loci.

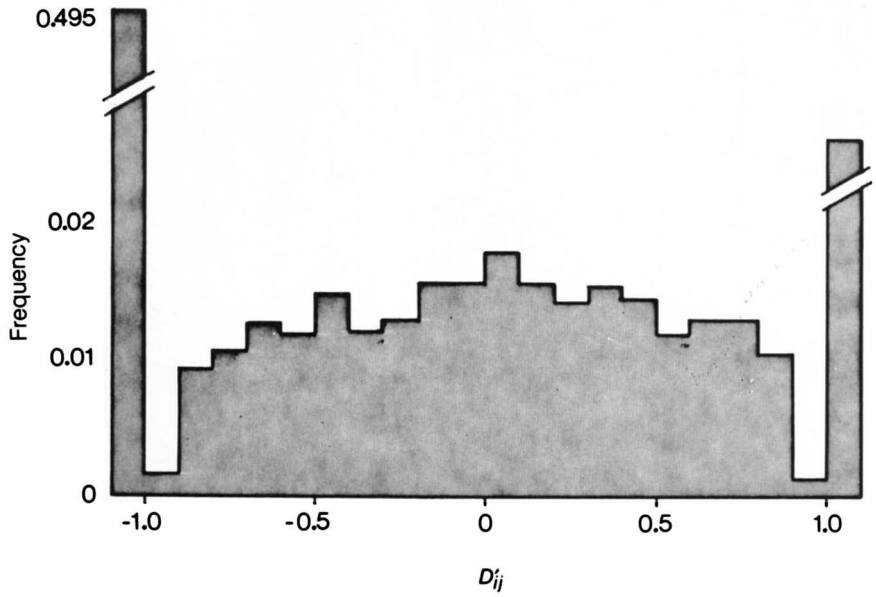

FIGURE 3.—The distribution of D'_{ij} in a sample of size $n = 200$ and $4Nc = 10$ when there are three alleles at one locus and six at the other.

Figure 4 gives the proportion of D'_{ij} values for which $D' = -1.0$, 1.0 or is between these values for $2n = 200$ and six alleles at each locus. When there is tight linkage, both the proportions in the $D' = -1.0$ and $D' = 1.0$ categories are elevated. As recombination becomes larger, the proportion in these two classes declines, and the proportion not having values of -1.0 or 1.0 greatly increases. For the example in Figure 4, the proportion not having values of -1.0 or 1.0 increases from 0.06 to 0.41 with increasing recombination, reflecting the presence of many of the gametic categories due to increased recombination. The proportion in these categories is relatively constant for a given number of alleles for different sample sizes, with a slight reduction in the $D' = -1.0$ and $D' = 1.0$ classes as sample size increases.

The proportion of D'_{ij} values in the -1.0 and 1.0 classes is dependent upon the number of alleles in the sample. Figure 5 gives the proportion in these three classes for different numbers of alleles when $n = 200$ and $4Nc = 10$. As the number of alleles increases, the proportion of values in the $D' = -1.0$ class increases and that in the $D' = 1.0$ decreases. The relative size of these values reflects the conditions under which $D' = -1.0$ and $D' = 1.0$ values are attained. For $D'_{ij} = -1.0$, either A_iB_j or $\overline{A}_i\overline{B}_j$ must be absent from the sample where $\overline{A}_i$ and $\overline{B}_j$ indicate all alleles at the A and B loci except A_i and B_j, respectively. Generally, with multiple alleles the first case occurs, *i.e.*, $x_{ij} = 0$. On the other hand, for $D_{ij} = 1$, x_{ij} must be either p_i or q_j, *i.e.*, either $A_i\overline{B}_j$ or $\overline{A}_iB_j$ is missing. Although there are two ways in which D'_{ij} may be 1, with multiple alleles both are much less likely to occur than to obtain a sample with $D'_{ij} = -1$. For example, in a sample of size n, and assuming that the population disequilibrium value is $\overline{D}_{ij} = 0$, the probability of $D'_{ij} = -1$ is

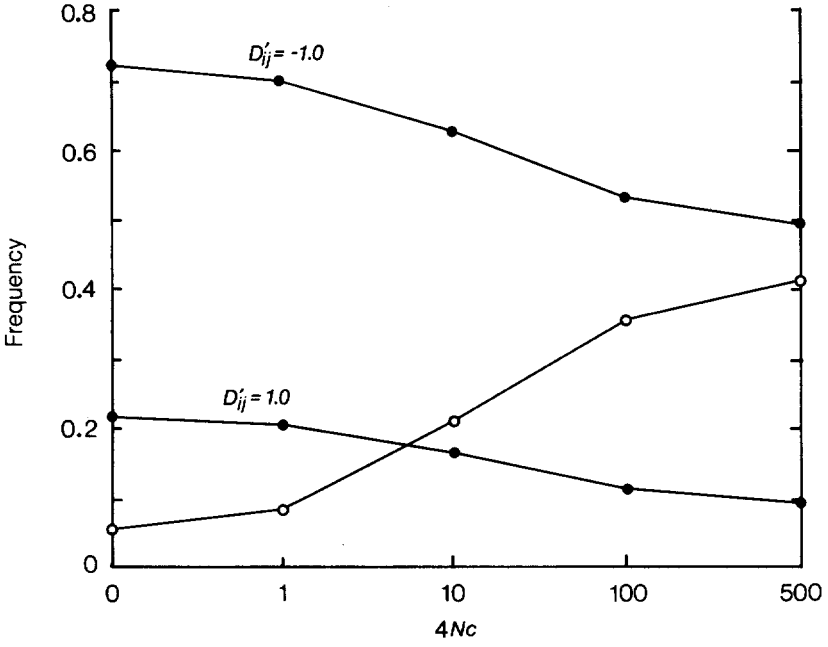

FIGURE 4.—The frequency of D'_{ij} values that are −1.0, 1.0 (closed circles) and intermediate (open circles) for different levels of recombination when n = 200 and there are six alleles at both loci.

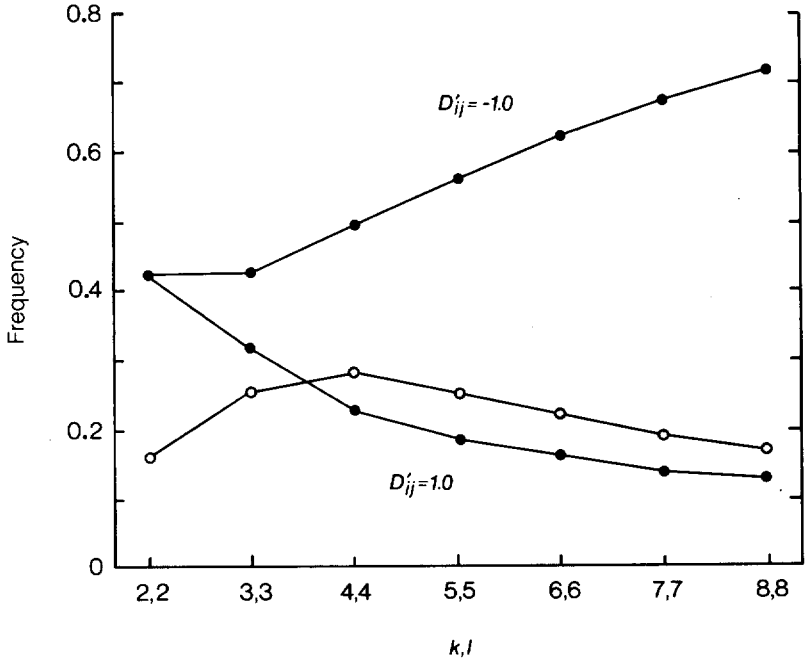

FIGURE 5.—The frequency of D'_{ij} values that are −1.0, 1.0 (closed circles) and intermediate (open circles) for different numbers of alleles when n = 200 and $4Nc$ = 10.

$$P(D'_{ij} = -1 \mid \overline{D}_{ij} = 0) = (1 - p_iq_j)^n + [1 - (1 - p_i)(1 - q_j)]^n. \quad (10)$$

The probability of $D'_{ij} = 1$ is

$$P(D'_{ij} = 1 \mid \overline{D}_{ij} = 0) = (1 - p_i)^n + (1 - q_j)^n. \quad (11)$$

Assuming that for multiple alleles p_i and q_j are much less than 0.5, then (10) is much greater than (11), explaining the larger proportion of $D'_{ij} = -1$ as compared to $D'_{ij} = 1$.

Tables 1, 2 and 3 give the average simulated values for the three measures of the total disequilibrium, D^*, F^* and Q^*, as given in expressions (5), (7) and (9), respectively. In all cases, the means were calculated as the average value over all replicate samples. The 95% intervals were obtained by ordering the disequilibrium values for the replicate samples for a given k, l, n and $4Nc$ combination and then determining the values at the 2.5% and 97.5% points of this distribution.

The three measures are all highly dependent on the level of recombination for given k, l and n values and decrease as recombination increases. In other words all three measures approach zero as $4Nc$ increases (actually Q^* approaches $1/n$). Note that F^* is near zero even when $4Nc = 100$. The magnitude of D^* and F^* is relatively constant as sample size increases conditioned on given k, l and $4Nc$ values. Q^*, on the other hand, is sample size dependent, decreasing with an increase in sample size. As expected when $4Nc$ is large, Q^* decreases inversely with increases in sample size.

The pattern of disequilibrium values for different numbers of alleles is different for the three measures. D^* is low for low k, l values; increases as k, l increases; and decreases as k and l become large for low $4Nc$ values. F^* increases with an increase in allelic numbers, whereas Q^* is at a maximum at low k, l values and decreases as k, l increases. This property of Q^* makes it useful for extrapolation to higher k and l values, because its expectation always decreases.

The confidence limits of the disequilibrium measures are generally quite large. For low $4Nc$ values, the 95% intervals extend throughout or nearly throughout the range of possible values. This occurs because the probability of high disequilibrium even under neutrality is often substantial. However, as $4Nc$ increases, the confidence limits become narrower. As with the means, the intervals for D^* and F^* are relatively similar for different sample sizes, and the upper limit for Q^* decreases with increasing sample size. The upper limits of D^* and Q^* are high for low number of alleles because the maximum value for these measures is dependent upon allele number. On the other hand, the upper limit for F^* is similar for different numbers of alleles.

Notice that the lower limit for F^* is slightly negative. Let us examine briefly a two-allele model to determine when these values occur. For two alleles,

$$\begin{aligned} F_{AB} &= (p_1q_1 + D)^2 + (p_1q_2 - D)^2 + (p_2q_1 - D)^2 + (p_2q_2 + D)^2 \\ &= p_1^2q_1^2 + p_1^2q_2^2 + p_2^2q_1^2 + p_2^2q_2^2 + 4D^2 + 2D(p_1q_1 - p_1q_2 - p_2q_1 + p_2q_2) \end{aligned}$$

TABLE 1

The value of D^*, as given in expression (5), for different sample sizes, numbers of alleles and recombination, with 95% intervals in parentheses

		$4Nc$				
n	k,l	0	1	10	100	500
100	2,2	0.164 (0.000, 1.000)	0.124 (0.000, 1.000)	0.045 (0.000, 0.412)	0.014 (0.000, 0.094)	0.011 (0.000, 0.074)
	2,3	0.206 (0.000, 0.927)	0.139 (0.000, 0.884)	0.051 (0.000, 0.363)	0.017 (0.000, 0.103)	0.012 (0.000, 0.067)
	3,3	0.221 (0.001, 0.872)	0.164 (0.001, 0.780)	0.065 (0.001, 0.341)	0.021 (0.001, 0.093)	0.012 (0.000, 0.048)
	4,4	0.228 (0.006, 0.698)	0.180 (0.006, 0.586)	0.066 (0.004, 0.245)	0.018 (0.002, 0.057)	0.012 (0.002, 0.041)
	6,6	0.205 (0.019, 0.450)	0.164 (0.021, 0.389)	0.073 (0.012, 0.189)	0.019 (0.005, 0.049)	0.012 (0.004, 0.030)
	8,8	0.157 (0.029, 0.319)	0.142 (0.032, 0.267)	0.069 (0.019, 0.143)	0.020 (0.007, 0.045)	0.012 (0.005, 0.024)
200	2,2	0.124 (0.000, 1.000)	0.093 (0.000, 1.000)	0.037 (0.000, 0.355)	0.013 (0.000, 0.073)	0.006 (0.000, 0.037)
	2,3	0.156 (0.000, 0.955)	0.113 (0.000, 0.809)	0.043 (0.000, 0.331)	0.011 (0.000, 0.059)	0.007 (0.000, 0.030)
	3,3	0.210 (0.000, 0.905)	0.149 (0.000, 0.818)	0.053 (0.000, 0.299)	0.013 (0.000, 0.061)	0.007 (0.000, 0.030)
	4,4	0.210 (0.002, 0.793)	0.157 (0.002, 0.576)	0.062 (0.002, 0.258)	0.013 (0.001, 0.050)	0.007 (0.001, 0.024)
	6,6	0.206 (0.009, 0.547)	0.163 (0.009, 0.431)	0.068 (0.008, 0.218)	0.014 (0.003, 0.038)	0.007 (0.002, 0.019)
	8,8	0.179 (0.027, 0.358)	0.155 (0.020, 0.355)	0.067 (0.014, 0.172)	0.013 (0.004, 0.032)	0.007 (0.003, 0.016)
400	2,2	0.103 (0.001, 1.000)	0.069 (0.000, 0.959)	0.023 (0.000, 0.245)	0.007 (0.000, 0.051)	0.003 (0.000, 0.020)
	2,3	0.122 (0.000, 0.969)	0.101 (0.000, 0.878)	0.033 (0.000, 0.262)	0.009 (0.000, 0.044)	0.004 (0.000, 0.023)
	3,3	0.166 (0.000, 0.911)	0.116 (0.000, 0.702)	0.044 (0.000, 0.276)	0.010 (0.000, 0.050)	0.004 (0.000, 0.018)
	4,4	0.222 (0.001, 0.834)	0.152 (0.001, 0.685)	0.051 (0.001, 0.228)	0.011 (0.000, 0.048)	0.004 (0.000, 0.017)
	6,6	0.214 (0.005, 0.573)	0.156 (0.007, 0.432)	0.063 (0.005, 0.202)	0.011 (0.002, 0.035)	0.004 (0.001, 0.013)
	8,8	0.184 (0.015, 0.432)	0.154 (0.013, 0.370)	0.066 (0.010, 0.169)	0.011 (0.003, 0.028)	0.005 (0.001, 0.012)

TABLE 2

The value of F^*, the identity excess, as given in expression (7), for different sample sizes, numbers of alleles and recombination, with 95% intervals in parentheses

		$4Nc$				
n	k,l	0	1	10	100	500
100	2,2	0.027 (−0.018, 0.249)	0.018 (−0.019, 0.237)	0.005 (−0.017, 0.098)	0.001 (−0.014, 0.026)	0.001 (−0.010, 0.022)
	2,3	0.042 (−0.022, 0.236)	0.026 (−0.022, 0.217)	0.007 (−0.021, 0.091)	0.001 (−0.016, 0.030)	0.001 (−0.014, 0.024)
	3,3	0.055 (−0.028, 0.233)	0.038 (−0.027, 0.213)	0.013 (−0.024, 0.101)	0.003 (−0.019, 0.040)	0.001 (−0.017, 0.025)
	4,4	0.074 (−0.031, 0.220)	0.058 (−0.032, 0.197)	0.018 (−0.028, 0.098)	0.004 (−0.019, 0.035)	0.002 (−0.020, 0.026)
	6,6	0.096 (−0.018, 0.202)	0.075 (−0.020, 0.188)	0.031 (−0.025, 0.101)	0.007 (−0.018, 0.035)	0.004 (−0.016, 0.027)
	8,8	0.090 (−0.006, 0.171)	0.079 (−0.013, 0.161)	0.035 (−0.009, 0.082)	0.009 (−0.014, 0.034)	0.005 (−0.012, 0.020)
200	2,2	0.021 (−0.017, 0.248)	0.014 (−0.015, 0.201)	0.003 (−0.014, 0.067)	0.001 (−0.010, 0.022)	0.000 (−0.009, 0.012)
	2,3	0.030 (−0.023, 0.238)	0.021 (−0.020, 0.203)	0.005 (−0.019, 0.075)	0.001 (−0.013, 0.024)	0.000 (−0.010, 0.015)
	3,3	0.050 (−0.022, 0.237)	0.033 (−0.025, 0.212)	0.009 (−0.023, 0.090)	0.002 (−0.015, 0.027)	0.001 (−0.012, 0.016)
	4,4	0.062 (−0.028, 0.222)	0.048 (−0.029, 0.202)	0.013 (−0.029, 0.094)	0.002 (−0.018, 0.034)	0.001 (−0.014, 0.020)
	6,6	0.084 (−0.031, 0.203)	0.068 (−0.027, 0.178)	0.025 (−0.022, 0.095)	0.004 (−0.017, 0.029)	0.002 (−0.014, 0.019)
	8,8	0.092 (−0.010, 0.182)	0.079 (−0.015, 0.180)	0.030 (−0.021, 0.086)	0.005 (−0.015, 0.026)	0.003 (−0.011, 0.019)
400	2,2	0.017 (−0.017, 0.243)	0.010 (−0.015, 0.158)	0.002 (−0.014, 0.045)	0.000 (−0.009, 0.015)	0.000 (−0.006, 0.008)
	2,3	0.022 (−0.019, 0.238)	0.017 (−0.020, 0.193)	0.004 (−0.018, 0.071)	0.000 (−0.011, 0.019)	0.000 (−0.007, 0.012)
	3,3	0.037 (−0.023, 0.237)	0.024 (−0.022, 0.193)	0.006 (−0.021, 0.080)	0.001 (−0.015, 0.023)	0.000 (−0.009, 0.012)
	4,4	0.063 (−0.027, 0.231)	0.042 (−0.026, 0.197)	0.010 (−0.026, 0.088)	0.001 (−0.016, 0.025)	0.001 (−0.011, 0.015)
	6,6	0.084 (−0.023, 0.215)	0.060 (−0.028, 0.181)	0.022 (−0.025, 0.096)	0.004 (−0.016, 0.029)	0.001 (−0.011, 0.015)
	8,8	0.088 (−0.019, 0.193)	0.071 (−0.023, 0.186)	0.029 (−0.018, 0.090)	0.004 (−0.017, 0.023)	0.002 (−0.010, 0.015)

TABLE 3

The value of Q^*, as given in expression (9), for different sample sizes, numbers of alleles and recombination, with 95% intervals in parentheses

		$4Nc$				
n	k,l	0	1	10	100	500
100	2,2	0.164 (0.000, 1.000)	0.124 (0.000, 1.000)	0.045 (0.000, 0.412)	0.014 (0.000, 0.094)	0.011 (0.000, 0.074)
	2,3	0.148 (0.000, 0.500)	0.101 (0.000, 0.500)	0.042 (0.000, 0.303)	0.016 (0.000, 0.088)	0.012 (0.000, 0.062)
	3,3	0.117 (0.005, 0.378)	0.091 (0.000, 0.290)	0.042 (0.000, 0.189)	0.018 (0.000, 0.087)	0.011 (0.000, 0.048)
	4,4	0.093 (0.002, 0.224)	0.080 (0.002, 0.222)	0.039 (0.001, 0.128)	0.016 (0.000, 0.063)	0.011 (0.001, 0.030)
	6,6	0.070 (0.008, 0.125)	0.061 (0.010, 0.122)	0.035 (0.005, 0.078)	0.016 (0.003, 0.046)	0.011 (0.003, 0.033)
	8,8	0.054 (0.020, 0.090)	0.050 (0.016, 0.085)	0.032 (0.011, 0.056)	0.016 (0.005, 0.034)	0.011 (0.004, 0.026)
200	2,2	0.124 (0.000, 1.000)	0.093 (0.000, 1.000)	0.037 (0.000, 0.355)	0.013 (0.000, 0.073)	0.006 (0.000, 0.037)
	2,3	0.111 (0.000, 0.500)	0.085 (0.000, 0.500)	0.034 (0.000, 0.260)	0.011 (0.000, 0.063)	0.007 (0.000, 0.030)
	3,3	0.109 (0.000, 0.280)	0.072 (0.000, 0.258)	0.031 (0.000, 0.182)	0.010 (0.000, 0.046)	0.006 (0.000, 0.029)
	4,4	0.075 (0.000, 0.222)	0.063 (0.000, 0.205)	0.030 (0.001, 0.122)	0.011 (0.000, 0.052)	0.006 (0.000, 0.020)
	6,6	0.056 (0.003, 0.120)	0.045 (0.002, 0.115)	0.027 (0.003, 0.070)	0.010 (0.002, 0.031)	0.006 (0.001, 0.020)
	8,8	0.047 (0.012, 0.082)	0.040 (0.008, 0.075)	0.024 (0.005, 0.053)	0.009 (0.003, 0.024)	0.006 (0.002, 0.018)
400	2,2	0.103 (0.000, 1.000)	0.069 (0.000, 0.959)	0.023 (0.000, 0.245)	0.007 (0.000, 0.051)	0.003 (0.000, 0.020)
	2,3	0.080 (0.000, 0.500)	0.069 (0.000, 0.500)	0.024 (0.000, 0.188)	0.008 (0.000, 0.038)	0.003 (0.000, 0.018)
	3,3	0.079 (0.000, 0.270)	0.057 (0.000, 0.253)	0.024 (0.000, 0.149)	0.008 (0.000, 0.042)	0.004 (0.000, 0.016)
	4,4	0.066 (0.000, 0.222)	0.052 (0.000, 0.181)	0.024 (0.000, 0.111)	0.007 (0.000, 0.033)	0.004 (0.000, 0.013)
	6,6	0.049 (0.001, 0.120)	0.040 (0.001, 0.095)	0.022 (0.001, 0.062)	0.008 (0.001, 0.030)	0.004 (0.001, 0.013)
	8,8	0.038 (0.005, 0.078)	0.033 (0.004, 0.067)	0.020 (0.003, 0.045)	0.007 (0.001, 0.023)	0.004 (0.001, 0.011)

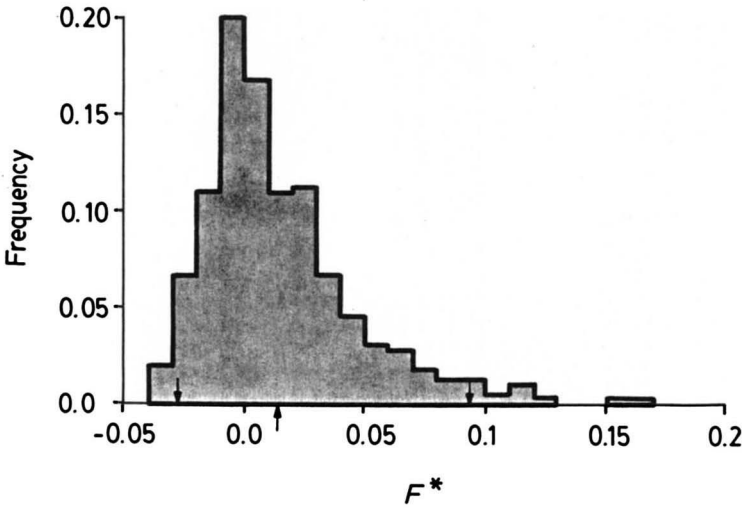

FIGURE 6.—The distribution of F^* values from 452 replicate samples for $n = 200$, $4Nc = 10$ and four alleles at each locus.

and

$$F_A F_B = p_1^2 q_1^2 + p_1^2 q_2^2 + p_2^2 q_1^2 + p_2^2 q_2^2$$

so that

$$F^* = 4D^2 + 2D(p_1 q_1 - p_1 q_2 - p_2 q_1 + p_2 q_2). \tag{12}$$

In other words, F^* is negative whenever the right-hand term is negative and greater than (absolutely) the left-hand term, a situation that occurs at a low frequency.

As an illustration of the distributions of the conditioned disequilibrium measures, the distributions of F^* and Q^* are given for two examples in Figures 6 and 7 (the distribution of D^* is quite similar to Q^*). Notice that for both figures there is a large tail to the right, indicating that high disequilibrium samples are quite possible. The mean is indicated by the upward-pointing arrow, and the 95% intervals are indicated by the pair of downward-pointing arrows.

Expression (5), D^*, as we have used it is different from

$$\sigma_D^2 = \frac{E(D^2)}{E[(1 - F_A)(1 - F_B)]} \tag{13}$$

as given by HILL (1975) and OHTA (1980). We calculated the mean of the ratio D^2 divided by the product of the heterozygosities, whereas they calculated the ratio of the expectations of these quantities. Obviously, the ratio of the expectations is more mathematically tractable, but the mean of the ratio allowed us to calculate 95% confidence limits. In addition, D^* is the experimentally observable quantity (MARUYAMA 1982).

As expected, D^* and σ_D^2 are not equivalent throughout the range of k, l, n and $4Nc$ values. Table 4 gives σ_D^2 for a number of parameter combinations, as

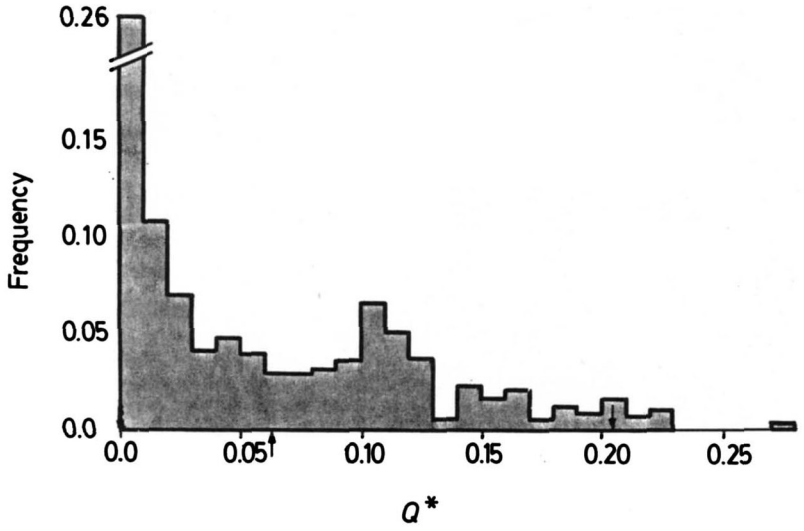

FIGURE 7.—The distribution of Q^* values for 423 replicate samples for $n = 200$, $4Nc = 1$ and four alleles at each locus.

TABLE 4

The mean value of σ_D^2, the standardized linkage disequilibrium, as given in expression (12), for several sample sizes, number of alleles and recombination values, with the ratio, σ_D^2/D^*, in parentheses

		$4Nc$		
n	k,l	0	10	500
100	2,2	0.454 (2.76)	0.116 (2.57)	0.013 (1.19)
	4,4	0.299 (1.31)	0.080 (1.21)	0.013 (1.04)
	8,8	0.165 (1.05)	0.071 (1.03)	0.012 (1.10)
400	2,2	0.482 (4.68)	0.068 (2.95)	0.004 (1.24)
	4,4	0.344 (1.55)	0.074 (1.46)	0.005 (1.06)
	8,8	0.201 (1.10)	0.071 (1.08)	0.005 (1.01)

well as the ratio of σ_D^2 over D^*. σ_D^2 is much greater than D^* for low values of k and l, particularly when $4Nc$ is small, whereas the two values are similar when $4Nc$ is large. More specifically, unlike D^*, σ_D^2 is quite dependent on allele number in the sample when $4Nc$ is low; however, as allele number increases, σ_D^2 and D^* are much closer.

APPLICATIONS

As illustrations of the usefulness of these theoretical results, we will consider data from four different data sets: two from the major histocompatibility system in humans (HLA) and one each from electrophoretic loci in *E. coli* and lodgepole pine. Although we realize that probably none of these data sets come from populations or species that completely conform to a theoretical population, *i.e.*, are in equilibrium with long-term values of $4Nu$ and $4Nc$, comparison

TABLE 5

The observed values for three total disequilibrium measures for the *HLA-A* and *HLA-B* loci in samples from three South American Indian populations (from BLACK and SALZANO 1981) and the Hutterites (MORGAN *et al.* 1980)

	Parak	Tiriyo	Waiapi	Hutterites
n	196	196	234	406
k,l	4, 4	4, 5	5, 5	7, 13
D^*	0.229	0.120	0.0667	0.0566
F^*	0.126	0.0343	0.0288	0.0426
Q^*	0.099	0.076	0.065	0.037

to equilibrium predictions can give some insight into the important factors affecting disequilibrium in these species (see the discussions below).

HLA: The human histocompatibility system was initially examined in detail because of the necessity to match donors and recipients in tissue transplantation. This complex is composed of a large number of genes and is located on chromosome 6. The most well-defined loci are *HLA-A* and *HLA-B*, two genes that determine antigenic factors residing on the surface of many cell types. These loci are 0.8 map units apart, and each is highly variable, with large numbers of alleles in nearly every sample that has been examined (see THOMSON 1981 for a review). We should note that HLA data from a number of samples do not fit single-locus neutrality expectations (HEDRICK and THOMSON 1983; HEDRICK 1983a).

HLA-A and *HLA-B* data from three groups of South American Indians (BLACK and SALZANO 1981) and Hutterites from Canada (MORGAN *et al.* 1980) were analyzed. Table 5 gives the observed D^*, F^* and Q^* values for these samples. If these observed values are compared with those in Tables 1, 2 and 3 for the same sample size and numbers of alleles, the observed values suggest that $4Nc$ must be quite small. In fact, the observed values of all three measures in the Parak are greater than the simulated mean for $4Nc = 0$ (assuming $n = 200$, $k = l = 4$). The observed values fall within the 95% interval for D^* and Q^* for $4Nc = 10$. The observed F^*, on the other hand, is greater than the upper confidence limit for $4Nc = 10$. Notice Figure 6, which gives the conditional distribution of F^* in this case ($4Nc = 0$), and Figure 7, which gives the distribution of Q^* for $4Nc = 1$. The disequilibrium values for the Tiriyo, Waiapi and Hutterite samples are somewhat lower. For example, conservative comparisons of the Hutterite values can be made to Q^* with $n = 400$ and $k = l = 8$, remembering that Q^* declines with increasing numbers of alleles. In this case, the observed value lies between the mean for $4Nc = 0$ and 1.

If we assume the estimated recombination rate between these loci, $c = 0.008$, then for $4Nc = 1$, N would be 31 ($N = 312$ for $4Nc = 10$), a quite small effective population size. However, LI, NEEL and ROTHMAN (1978) have estimated the variance effective population size in a group of four Yanomama villages to be approximately 200, suggesting that there is a small finite population size in groups of South American Indians. Even so, there appears to be

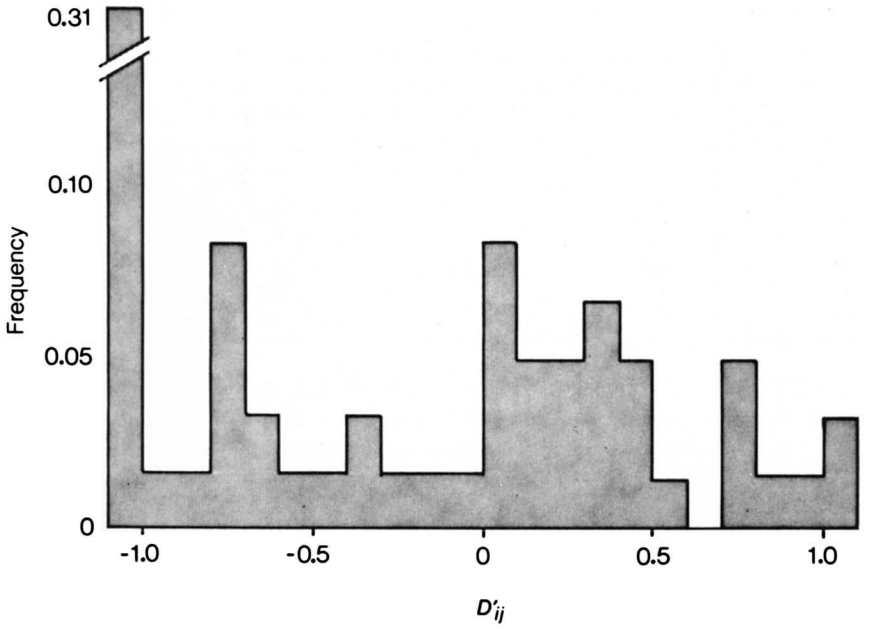

FIGURE 8.—The distribution of D'_{ij} values for two HLA loci in samples from three South American Indian populations (from BLACK and SALZANO 1981).

greater total disequilibrium among alleles at these loci than predicted for a sample taken from a neutrality population.

Some insight as to the type of deviation from neutrality expectation observed in these samples is possible by examining the distribution of D'_{ij} values. Figures 8 and 9 give these distributions for the three Indian samples and the Hutterite example, respectively. These observed distributions can be visually compared to the neutrality expectations (see Figure 2 for a typical example when $4Nc$ is large, and see Figure 3 for a smaller value of $4Nc$). Notice that the South American Indian values are spread out over the whole range of D' values, are somewhat asymmetrical and have few $D' = 1.0$ values. On the other hand, the distribution for the Hutterites has a high preponderance of $D' = -1.0$ values and has virtually no other negative D' values.

***E. coli*:** Electrophoretic analysis of enzyme variation in *E. coli* isolated from natural sources has revealed an extraordinary amount of allelic variation among haploid strains (*e.g.*, SELANDER and LEVIN 1980; WHITTAM, OCHMAN and SELANDER 1983b). In contrast to the HLA data, single locus genetic diversity values from a large sample of *E. coli* isolates have been shown to fit neutrality expectations (WHITTAM, OCHMAN and SELANDER 1983a). Because of the rarity of sexual reproduction in *E. coli*, one might expect disequilibrium to be near a neutrality maximum, *i.e.*, equivalent to $4Nc = 0$.

We have calculated the measures of total disequilibrium for a sample of 100 strains obtained from 100 Swedish schoolgirls (CAUGANT *et al.* 1983). Ten loci with various numbers of alleles have been used, giving 45 pairs of loci. The

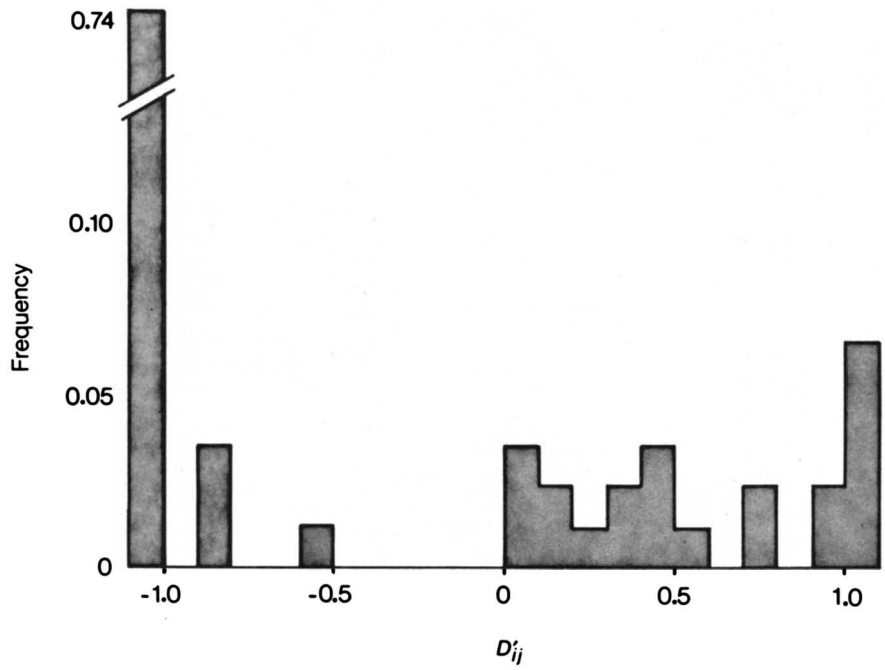

FIGURE 9.—The distribution of D' values for two HLA loci in the sample of Hutterites (from MORGAN *et al.* 1980).

average Q^* values are given in Figure 10, along with the number of pairs of loci with specific numbers of alleles (in parentheses). Surprisingly, these are all much less than that expected for $4Nc = 0$, although they are still within the 95% interval. Also given in Figure 10 are the theoretical values for $4Nc = 10$ and 100 for $n = 100$ that bracket the observed values except for five values, two of which are less than expected for $4Nc = 100$, and three of which are greater than expected for $4Nc = 10$. Figure 11 gives as an example the distribution of D'_{ij} values for the six loci pairs with three alleles at one locus and six alleles at the other. This distribution is quite similar to the neutrality distribution given in Figure 3 for $4Nc = 10$.

These total disequilibrium statistics suggest that either c is not zero, N is quite large or some other factor is important in effecting disequilibrium. However, it has been suggested that the rate of recombination for *E. coli* is quite low (*e.g.*, SELANDER and LEVIN 1980), approximately at the level of the mutation rate of 10^{-7}. A large effective population size is suggested by the large single-locus genetic diversity estimates (see the discussions in MARUYAMA and KIMURA 1980), although population subdivision is suggested by WHITTAM, OCHMAN and SELANDER (1983b). A factor that appears important in *E. coli* is genetic hitchhiking (*e.g.*, LEVIN 1981; HEDRICK 1982), although it appears from preliminary calculations that genetic hitchhiking will generally increase conditional disequilibrium.

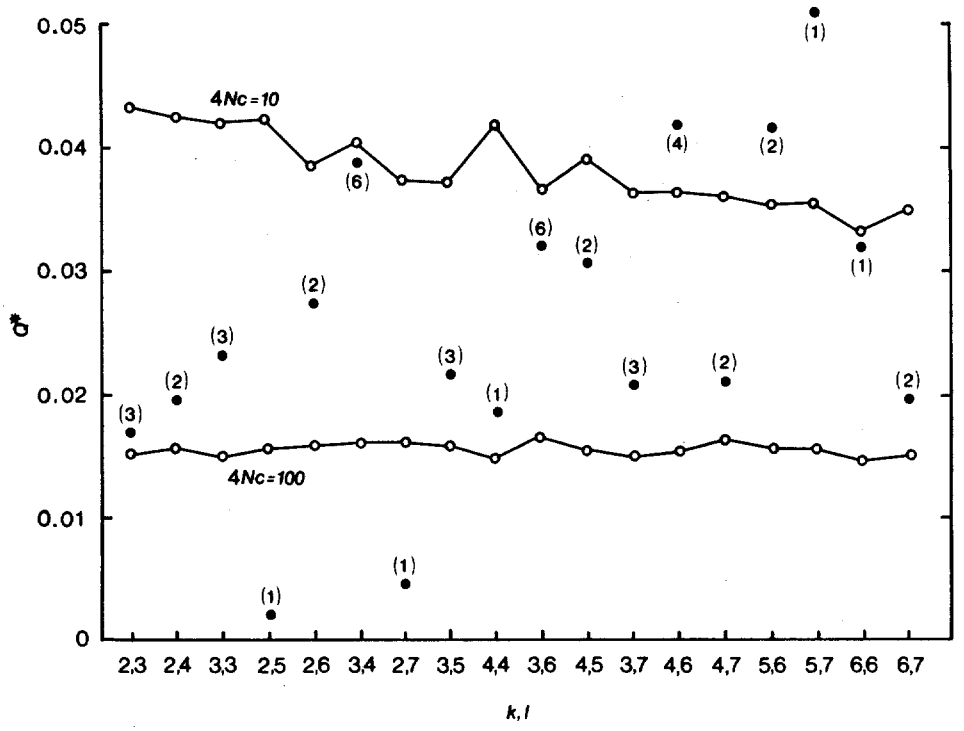

FIGURE 10.—Mean values of Q^* for 45 pairs of loci in *E. coli* (from CAUGANT *et al.* 1983), where the numbers in parentheses indicate the number of locus pairs for that k, l combination and the theoretical values for $4Nc = 10$ and 100.

Lodgepole pine: Except for a few higher organisms, such as *Drosophila melanogaster, Mus musculus* and humans, the map distance between linked loci often is not known precisely. One exception is for a group of electrophoretic loci in lodgepole pine, four of which are closely linked on one chromosome and two of which are linked on another chromosome (CONKLE 1981; EPPERSON 1983). EPPERSON (1983) surveyed the frequencies of alleles at these loci and the frequencies of two-locus gametes in samples of approximately $n = 400$ in two different populations. Here, we used the means of the most closely linked loci ($c \leqq 0.005$), each having three alleles. Even for these tightly linked pairs, disequilibrium values were quite low (Table 6), about the average expected value for $4Nc = 500$ in Tables 1, 2 and 3 when $n = 400$, $k = l = 3$. Using the estimated c values of 0.002 and 0.005 and neutrality equilibrium, then the effective population size would be 62,500 and 25,000, respectively, of the same magnitude suggested by EPPERSON (1983) based on other information. The distributions of D' values for the three loci $c < 0.002$ and $c = 0.005$, are given in Figure 12. This distribution has the same general pattern as found for neutrality, although there are many more values around $D' = 0$ and fewer $D' = -1$ and $D' = 1$ values (see Figure 2 for the pattern for neutrality when $n = 200$ and six alleles at both loci).

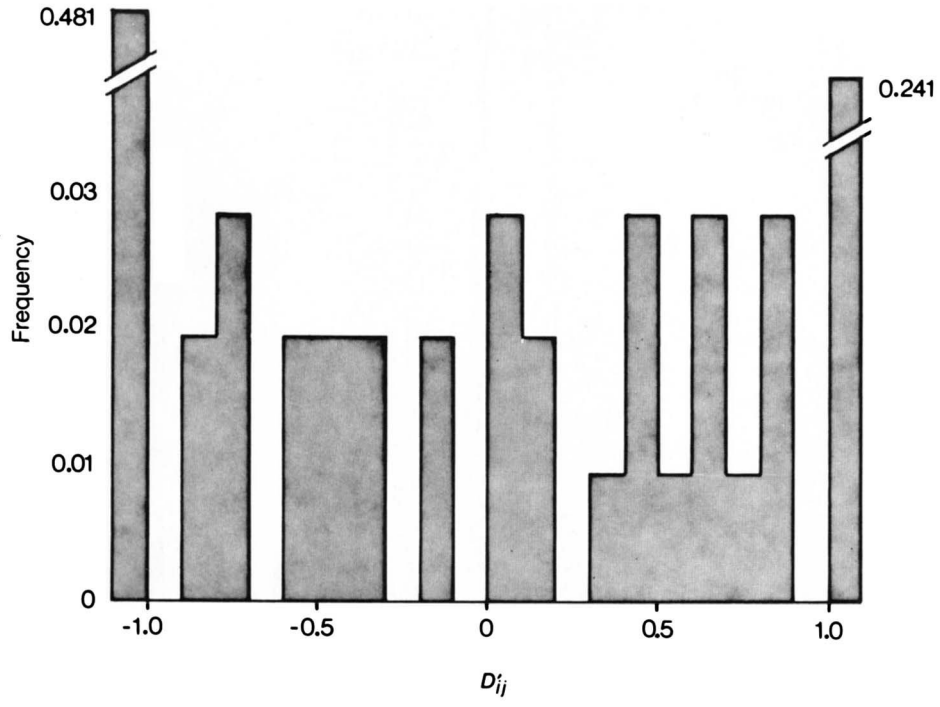

FIGURE 11.—The distribution of D'_{ij} for the six locus pairs in the *E. coli* data, in which there are three alleles at one locus and six at the other.

TABLE 6

The observed values for total disequilibrium measures for three electrophoretic locus pairs in samples of lodgepole pine from two locations (from EPPERSON 1983)

	Scar Mountain	Indian Creek
n	402	390
k,l	3, 3	3, 3
D^*	0.0048	0.00069
F^*	−0.0049	−0.00029
Q^*	0.0048	0.0021

DISCUSSION

Using computer simulation, we have given the expected disequilibrium and the 95% interval at a pair of neutral loci for a given sample size and number of alleles in the sample (Tables 1–3). These results are an extension to two loci of the single-locus theory of EWENS (1972) and WATTERSON (1978a,b). As expected, the total disequilibrium is highly dependent upon the rate of recombination in the population from which the sample is drawn. Similar to EWENS' results for homozygosity, the disequilibrium is not strongly dependent upon sample size. Furthermore, WATTERSON (1978b) showed that the conditional

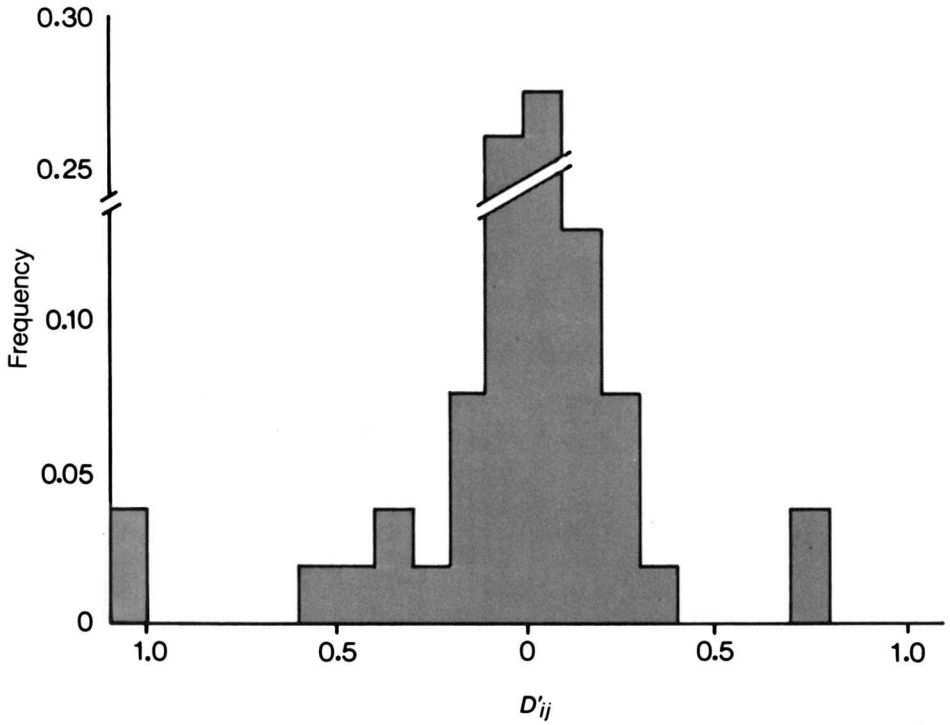

FIGURE 12.—The distribution of D'_{ij} for the three most closely linked loci in the lodgepole pine data (from EPPERSON 1983).

homozygosity is strongly dependent upon the number of alleles in the sample (the more alleles, the lower the homozygosity). We found that disequilibrium is dependent upon the number of alleles in the sample, although the pattern is dependent upon the measure of disequilibrium. For F^* the values increase with an increasing number of alleles, whereas for Q^* they decrease with increasing numbers of alleles. The 95% intervals are large, but become narrower for larger numbers of alleles and higher $4Nc$ values. Such wide intervals were suggested by the unconditional distribution generated by MARUYAMA (1982) and the $k = l = 2$ distribution given by HUDSON (1985). The bimodality at zero and unity that they both observed is typical when there are low numbers of alleles at both loci and low $4Nc$ values.

These expected values and their 95% intervals, as illustrated by the examples above, are useful in interpreting disequilibrium data. First, the disequilibrium between the *HLA-A* and *HLA-B* loci is higher than neutrality expectations, given the estimated recombination between these loci. Even allowing for a recent bottleneck and small population size, other factors, such as selection, probably are important. Second, the disequilibrium between pairs of electrophoretic loci in *E. coli* is consistent with neutrality expectations, but for relatively high $4Nc$ values, not for $4Nc = 0$. In this case, perhaps N is very large or there is enough sexual reproduction to make $4Nc > 0$. Last, the lodgepole

pine data are consistent with neutrality expectations, but for very large $4Nc$ values. Here, the observed disequilibrium is on the low end, even when the estimated c values are used and very large population sizes are assumed.

What is the expected influence on conditional disequilibrium of various evolutionary factors? These effects have not been explored to the same extent as those for single-locus deviations (see HEDRICK and THOMSON 1983 for a review). It appears that a recent bottleneck or genetic drift would generally increase disequilibrium although the exact impact conditioned on the sample size and number of alleles in a sample is not intuitive. G. THOMSON and W. KLITZ (unpublished results) have shown that selection favoring particular gametic types results in a distribution of D'_{ij} values quite different from neutrality expectations. From other studies of gene flow, genetic hitchhiking and multilocus selection, it is known that these factors, particularly for biallelic loci, can increase disequilibrium (see HEDRICK 1983b for a review). However, their impact on overall disequilibrium for multiallelic loci in a sample of a given size with a given number of alleles is not obvious. It would be useful to know the specific conditions under which these factors affect disequilibrium and the differences from neutrality that they cause. Examination of the population genetic properties of tightly linked polymorphic loci is particularly pertinent at this moment, given recent advances that allow detailed genetic mapping of organisms using restriction fragment length polymorphisms. For example, disequilibrium patterns observed between loci may indicate the nature of evolutionary events that have affected a genetic region.

We thank R. HUDSON for the use of his elegant simulation program. We appreciate the comments of W. EWENS, W. HILL, R. HUDSON, W. KLITZ, E. LOUIS, B. WEIR, T. WHITTAM and several reviewers on various drafts of the manuscript and thank KAI LO for computer assistance. This work was supported by National Institutes of Health grant HD 12731.

LITERATURE CITED

BISHOP, Y. M. M., S. E. FEINBERG and P. W. HOLLAND, 1975 *Discrete Multivariate Analysis.* MIT Press, Cambridge.

BLACK, F. L. and F. M. SALZANO, 1981 Evidence for heterosis in the HLA system. Am. J. Hum. Genet. **33:** 894–899.

CAUGANT, D. A., B. R. LEVIN, G. LIDIN-JANISON, T. S. WHITTAM, C. SVANBORGEON and R. K. SELANDER, 1983 Genetic diversity and relationships among strains of *Escherichia coli* in the intestine and those causing urinary tract infections. Prog. Allergy **33:** 203–227.

CONKLE, M. T., 1981 Isozyme variation and linkage in six conifer species, pp. 11–17. In: *Isozymes of North American Forest Trees and Forest Insects,* Edited by M. T. CONKLE. Gen. Tech. Rep. PSW-48. Pacific Southwest Forest and Range Exp. Stn., Berkeley, California.

EPPERSON, B. K., 1983 Multilocus genetic structure of natural populations of lodgepole pine. Ph.D. Thesis, University of California at Davis.

EWENS, W. J., 1972 The sampling theory of selectively neutral alleles. Theor. Pop. Biol. **3:** 87–112.

EWENS, W. J., 1977 Population genetics theory in relation to the neutralist-selection controversy. Adv. Hum. Genet. **8:** 67–134.

GOLDING, G. B., 1984 The sampling distribution of linkage disequilibrium. Genetics **108:** 257–274.

HEDRICK, P. W., 1982 Genetic hitchhiking: a new factor in evolution? Bioscience **32:** 845–853.

HEDRICK, P. W., 1983a Neutrality or selection of HLA? Am. J. Hum. Genet. **35:** 1055–1057.

HEDRICK, P. W., 1983b *Genetics of Populations.* Jones and Bartlett, Boston.

HEDRICK, P. W., S. JAIN and L. HOLDEN,1978 Multilocus systems in evolution. Evol. Biol. **11:** 101–182.

HEDRICK, P. W. and G. THOMSON, 1983 Evidence for balancing selection at HLA. Genetics **104:** 449–456.

HILL, W. G., 1975 Linkage disequilibrium among multiple neutral alleles produced by mutation in finite population. Theor. Pop. Biol. **8:** 117–126.

HILL, W. G., 1981 Estimation of effective population size from data on linkage disequilibrium. Genet. Res. **38:** 209–216.

HUDSON, R. R., 1983 Properties of a neutral allele model with intragenic recombination. Theor. Pop. Biol. **23:** 183–201.

HUDSON, R. R., 1985 The sampling distribution of linkage disequilibrium under an infinite-allele model without selection. Genetics **109:** 611–631.

KIMURA, M. and T. OHTA, 1971 *Theoretical Aspects of Population Genetics.* Princeton University Press. Princeton, New Jersey.

LEVIN, B. R., 1981 Periodic selection, infectious gene exchange and the genetic structure of *E. coli* populations. Genetics **99:** 1–23.

LEWONTIN, R. C., 1964 The interaction of selection and linkage. I. General considerations; heterotic models. Genetics **49:** 49–67.

LI, H. F., J. V. NEEL and E. D. ROTHMAN, 1978 A second study of the survival of a neutral mutant in a simulated Amerindian population. Am. Nat. **112:** 83–96.

MARUYAMA, T., 1982 Stochastic integrals and their application to population genetics, pp. 151–166. In: *Molecular Evolution, Protein Polymorphism and the Neutral Theory,* Edited by M. KIMURA. Japan Scientific Societies Press, Tokyo.

MARUYAMA, T. and M. KIMURA, 1980 Genetic variability and effective population size when local extinction and recolonization of subpopulations are frequent. Proc. Natl. Acad. Sci. USA **77:** 6710–6714.

MORGAN, K., T. M. HOLMES, J. SCHLAUT, L. MARCHUK, T. KOVITHAVONGS, F. PAZDERKA and J. B. DOSSETOR, 1980 Genetic variability of HLA in the Dariusleut Hutterites. A comparative genetic analysis of the Hutterites, the Amish, and other selected Caucasian populations. Am. J. Hum. Genet. **32:** 246–257.

OHTA, T., 1980 Linkage disequilibrium between amino acids sites in immunoglobulin genes and other multigene families. Genet. Res. **36:** 181–197.

OHTA, T. and M. KIMURA, 1969 Linkage disequilibrium due to random genetic drift. Genet. Res. **13:** 47–53.

SELANDER, R. K. and B. R. LEVIN, 1980 Genetic diversity and structure in *Escherichia coli* populations. Science **210:** 545–547.

SOKAL, R. R. and F. J. ROHLF, 1981 *Biometry.* Ed. 2. W. H. FREEMAN, San Francisco.

TAKAHATA, N., 1982 Linkage disequilibrium, genetic distance, and evolutionary distance under a general model of linked genes or a part of the genome. Genet. Res. **39:** 63–77.

THOMSON, G., 1981 A review of theoretical aspects of HLA and disease associations. Theor. Pop. Biol. **20:** 168–208.

WATTERSON, G. A., 1978a An analysis of multi-allelic data. Genetics **88:** 171–179.

WATTERSON, G. A., 1978b The homozygosity test of neutrality. Genetics **88:** 405–417.

WHITTAM, T. S., H. OCHMAN and R. K. SELANDER, 1983a Multilocus genetic structure in natural populations of *Escherichia coli*. Proc. Natl. Acad. Sci. USA **80:** 1751–1755.

WHITTAM, T. S., H. OCHMAN and R. K. SELANDER, 1983b Geographic components of linkage disequilibrium in natural populations of *Escherichia coli*. Mol. Biol. Evol. **1:** 67–83.

Communicating editor: W. J. EWENS